АНАЛИТИКА И ИСПОЛЬЗОВАНИЕ КОЛИЧЕСТВЕННОГО АНАЛИЗА ДАННЫХ

Хорев О.Е.
Уральский федеральный университет имени первого Президента России Б.Н. Ельцина
проспект Мира, 19, Екатеринбург, Свердловская обл.,620002, Россия
тел.: (343) 375-45-10, e-mail: khorev.oleg@urfu.ru

Аннотация — Решения в коммерческих и некоммерческих организациях принимают исходя из целого ряда факторов: опыта, интуиции, результатов экспериментов, аналитических исследований и накопленных данных. В коммерческих организациях традиционная аналитика чаще всего применяется для поддержки внутренних решений компании: «Сколько должен стоить этот продукт?» или «Как стимулировать покупателей совершать у нас покупки?» Аналитика в среде больших данных часто используется для разработки новых видов продуктов или дополнительных потребительских свойств.

ANALYTICS AND USE OF THE QUANTITATIVE ANALYSIS OF DATA

Khorev.O.E.[1]
[1] Ural Federal University named after the first President of Russia B.N.Yeltsin
pr. Mira, 19, Yekaterinburg, Sverdlovsk region, 620002, Russian Federation
ph.: 375-45-10, e-mail: khorev.oleg@urfu.ru

Abstract — Decisions in commercial and non-profit organizations accept proceeding from a number of factors: experience, intuition, results of experiments, analytical researches and the saved-up data. In the commercial organizations the traditional analytics is most often applied to support of internal decisions of the company: "How many has to cost this product?" or "How to stimulate buyers to make at us purchases?" The analytics in the environment of Big Data is often used for development of new types of products or padding consumer properties.

I. Введение

Аналитикой называется всестороннее использование баз данных, статистический и количественный анализ, объяснительные и прогнозные модели, а также доказательный менеджмент, применяемые для поддержки решений и увеличения ценности для потребителей [4].

В зависимости от цели и методов аналитику обычно делят на следующие типы: описательная, предсказательная, нормативная.

Описательная аналитика ориентирована на информирование, а характеристиках данных. Обычно описывает настоящую ситуацию и не говорит о причинах или последствиях имеющейся ситуации. Данный вид аналитики включат в себя: сбор, систематизацию, представление данных в табличной форме и последующее выделение основных характеристик описываемой области.

Предсказательная аналитика уже не ограничивается описание данных, а позволяет прогнозировать динамику будущих показателей, основываясь на данных за прошлые периоды. Сначала определяются связи между переменными, а затем на основе их анализа оценивается вероятность того или иного события: например, насколько вероятно, что потребитель отреагирует на рекламу и купит данный продукт. Хотя связи между переменными используются для прогнозирования будущего, явная причинно-следственная связь обнаруживается далеко не всегда. По сути, она совсем не обязательна для получения точного прогноза.

Нормативная аналитика включает в себя проведение экспериментов и оптимизацию, и за счет этого она ориентируется на более широкий круг задач. Данных вид аналитики предлагает направление дальнейших действий. Эксперимент в данном виде аналитики помогает ответить на вопросы о причинах произошедших явлений и сделать выводы о причинных связях. Для этого исследователи изменяют одну или несколько независимых переменных и наблюдают реакцию зависимой переменной, одновременно контролируя внешние по отношению к исследуемой системе факторы. Оптимизация же направлена на выявление оптимального значения конкретной переменной во взаимосвязи с другой переменной.

II. Основная часть

ЭТАПЫ АНАЛИТИЧЕСКОГО ПОДХОДА

В основе аналитического подхода количественного анализа лежит три крупных этапа, которые включают в себя шесть важных шагов, необходимых для проведения анализа. Ниже представлены данные этапы и шаги, которые они в себя включают (по классификации Дэвенпорта):

- формулирование проблемы;
  a) определение проблемы;
  b) изучение предыдущих поисков решения.
- решение проблемы;
  a) моделирование ситуации;
  b) сбор данных;
  c) анализ данных.
- результаты и необходимые меры.
  a) демонстрация результатов.

ЭТАП 1 ФОРМУЛИРОВАНИЕ ПРОБЛЕМЫ

Количественный анализ начинается с идентификации проблемы и подходов к ее решению. В анализе принятия решений этот этап называется формулированием проблемы и считается одним из наиболее важных для получения оптимального решения. Информацию для формулирования проблемы можно получить разными способами:

• обыкновенное любопытство (здравый смысл, наблюдение за событиями);
• опыт работы;
• потребность в решении либо действии;

• актуальные события, требующие внимания (сотрудника, организации в целом, нации);
• ранее проводившиеся исследования и уже существующие концепции;
• разработка проектов решений и анализ имеющегося и необходимого финансирования.

Стоит отметить, что на данном этапе анализ практически отсутствует здесь формируются предположения, которые в будущем должны будут проверять аналитики с помощью средств и методов анализа. На данном этапе самым главным является осознание проблемы и четкое понимание того, от чего данная задача так актуальна.

После того как проблема определена, следует выяснить, проводились ли ее исследования ранее и каковы были их результаты. Это второй шаг первого этапа количественного анализа (формулирование проблемы), поскольку информация о предшествующих исследованиях помогает аналитику и менеджеру оценить разные варианты формулировки проблемы и ее концептуализации. Обзор предшествующих исследований помогает уточнить ряд вопросов:
• В чем особенности нашего исследовательского проекта? Включает ли он опрос, предсказание, эксперимент, отчет?
• Какие данные нам необходимо собрать?
• Какие параметры изучали в предшествующих исследованиях?
• Какие виды анализа нам придется провести?
• Будут ли результаты нашего анализа отличаться от полученных ранее и как представить их в интересной форме?

Одна из ключевых особенностей количественного анализа (и вообще научного метода исследований) – это учет результатов более ранних исследований. Например, поиск относящейся к теме информации в книгах, отчетах и статьях очень важен для всестороннего понимания проблемы. Это помогает установить ключевые параметры и связи между ними. Комплексный обзор результатов любых предыдущих исследований той же тематики обязателен для любого вида количественного анализа. В аналитике невозможно получить нечто из ничего. Приступать к решению проблемы можно только ознакомившись с опытом тех, кто делал это ранее. Одна только систематизация и оценка имеющейся информации играет важную роль в уточнении модели анализа или подходов к решению проблемы.

ЭТАП 2 РЕШЕНИЕ ПРОБЛЕМЫ

Первым шагом на данном этапе является моделирование данных. Модель - объект-заместитель объекта-оригинала, обеспечивающий изучение некоторых свойств оригинала [1]. Модель можно сравнить с карикатурой. Она заостряет внимание на некоторых чертах – носе, улыбке, кудрях, – и на их фоне другие черты теряют выразительность. Хорошая карикатура отличается тем, что отдельные черты выбираются обдуманно и эффективно. Точно так же модель акцентирует внимание на отдельных особенностях реального мира. При построении любой модели вам придется действовать избирательно. Нужно выбрать именно те особенности, которые имеют отношение к решению вашей проблемы, и пренебречь остальными. Модель носит схематичный характер, чтобы помочь пользователю сфокусироваться на исследуемой проблеме [6].

На этом этапе разработка модели требует логического мышления, опыта и знакомства с предшествующими исследованиями. Только в этом случае можно с большой долей уверенности предположить, какие зависимые (те, которые нужно прогнозировать или объяснить) или независимые факторы сыграют основную роль. Можно попытаться протестировать модель – именно это отличает аналитическое мышление от менее точных методов принятия решений вроде интуиции.

Следующим шагом данного этапа является сбор данных и измерения выбранных переменных. Измерение – это определение значения переменной; массив данных – это набор таких значений. Существуют разные способы измерения. Сформулированная проблема сначала представляется в виде набора переменных в процессе моделирования, а затем приобретает вид массива данных в результате измерения. Выделяют следующие способы измерения данных:

Двоичные переменные. Данных вид переменных может принимать лишь два значения: 0 и 1. Их лучше всего применять для отражения факта наличия или отсутствия какого-либо критерия.

Ординальные переменные. Они имеют упорядоченные количественные значения, которые изменяются в зависимости от усиления или ослабления выраженности признака.

Количественные переменные. Эти переменные выражены числами обычно в стандартных единицах и наиболее хорошо подходят для проведения различных видов анализа.

Сами же данные могут быть предоставлены для анализа другими людьми. Однако подобные данные все равно лучше проверить на наличие в них ошибок и неточностей. Некоторые данные могут быть взяты из проводимых ранее исследований. Но если же вам не были предоставлены данные, то их можно найти самому различными способами. Для самостоятельного поиска данных могут быть использованы следующие источники:

- Поисковые системы.
- Поиск информации у первоисточника.
- Базы данных университетов.
- Источники данных общего характера.
- Тематические данные (спорт, география, государственные организации и пр) [5].

Так же для сбора и извлечения данных возможно использование методов парсинга, проведения опросов, наблюдений, экспериментов и др.

Третьим шагом данного этапа является анализ данных. Данный термин является переводом термина Data Mining. Термин Data Mining получил свое название из двух понятий: поиска ценной информации в большой базе данных (data) и добычи горной руды (mining). Оба процесса требуют или просеивания огромного количества сырого материала, или разумного исследования и поиска искомых ценностей.

Data Mining – мультидисциплинарная область, возникшая и развивающаяся на базе таких наук как прикладная статистика, распознавание образов, искусственный интеллект, теория баз данных и др.

Data Mining - это процесс обнаружения в сырых данных ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации

знаний, необходимых для принятия решений в различных сферах человеческой деятельности.

Суть и цель технологии Data Mining можно охарактеризовать так: это технология, которая предназначена для поиска в больших объемах данных неочевидных, объективных и полезных на практике закономерностей [11].

Для проведения анализа в данной работе были описаны следующие методы: кластерный анализ, факторный анализ, уровень значимости, проверка гипотез, дисперсионный анализ. Рассмотрим кратко каждый из этих методов.

Кластерный анализ включает в себя набор различных алгоритмов классификации. Задача кластерного анализа состоит в разбиении исходной совокупности объектов на группы схожих, близких между собой объектов. Эти группы называют кластерами. Другими словами, кластерный анализ – это один из способов классификации объектов по их признакам. Желательно, чтобы результаты классификации имели содержательную интерпретацию. Кластерный анализ позволяет рассматривать достаточно большой объём информации и сильно сжимать большие массивы социально-экономической информации, делать их компактными и наглядными [9].

Факторный анализ – методика комплексного и системного изучения и измерения воздействия факторов на величину результативного показателя. Факторный анализ позволяет решить две важные проблемы исследователя: описать объект измерения всесторонне и в то же время компактно. С помощью факторного анализа возможно выявление скрытых переменных факторов, отвечающих за наличие линейных статистических корреляций между наблюдаемыми переменными.

Две основных цели факторного анализа:

- определение взаимосвязей между переменными, (классификация переменных), то есть «объективная R-классификация»;
- сокращение числа переменных необходимых для описания данных [2].

Уровнем значимости называется такое максимальное отношение количества нетипичных выборочных значений (выбросов) ко всему объему выборки, что нулевая гипотеза отклоняется [3]. Иными словами, уровень значимости показывает количество нетипичных наблюдений (выборочных значений), необходимых для признания ложности нулевой гипотезы. Обычно уровень значимости задается как 5 процентов (0,05), но в ситуациях, когда предъявляются особенно строгие требования к доказательству истинности альтернативной гипотезы, этот показатель может быть задан и на более низком уровне, например, 1 процент (0,01). Значение α, равное 5 процентам, означает, что для отбрасывания нулевой гипотезы как ложной достаточно наличия менее 5 процентов нетипичных данных от их общего количества (при условии истинности нулевой гипотезы). На практике это требование часто проверяется путем расчета p-значения. Если p-значение меньше, чем α, то нулевая гипотеза признается ложной, а альтернативная гипотеза – истинной.

Проверка статистической гипотезы — это процесс принятия решения о том, противоречит ли рассматриваемая статистическая гипотеза наблюдаемой выборке данных. Часто делают выборку, чтобы определить аргумен-ты против гипотезы относительно популяции (генеральной совокупности). Этот процесс известен как проверка гипотез (проверка статистических гипотез или проверка значимости), он представляет количественную меру аргументов про-тив определенной гипотезы.

Установлено 5 стадий при проверке гипотез:

1. Определение нулевой (h0) и альтернативной гипотезы (h1) при исследовании. Определение уровня значимости критерия. Отбор необходимых данных из выборки.
2. Вычисление значения статистики критерия, отвечающей h0.
3. Вычисление критической области, проверка статистики критерия на предмет попадания в критическую область.
4. Интерпретация достигнутого уровня значимости p и результатов [10].

Дисперсионный анализ применяется для исследования влияния одной или нескольких качественных переменных (факторов) на одну зависимую количественную переменную.

В основе дисперсионного анализа лежит предположение о том, что одни переменные могут рассматриваться как причины (факторы, независимые переменные, а другие как следствия (зависимые переменные). Независимые переменные называют иногда регулируемыми факторами именно потому, что в эксперименте исследователь имеет возможность варьировать ими и анализировать получающийся результат.

Основной целью дисперсионного анализа является исследование значимости различия между средними с помощью сравнения (анализа) дисперсий. Разделение общей дисперсии на несколько источников, позволяет сравнить дисперсию, вызванную различием между группами, с дисперсией, вызванной внутригрупповой изменчивостью. При истинности нулевой гипотезы (о равенстве средних в нескольких группах наблюдений, выбранных из генеральной совокупности), оценка дисперсии, связанной с внутригрупповой изменчивостью, должна быть близкой к оценке межгрупповой дисперсии.

Сущность дисперсионного анализа заключается в расчленении общей дисперсии изучаемого признака на отдельные компо-ненты, обусловленные влиянием конкретных факторов, и проверке гипотез о значимости влияния этих факторов на исследуемый признак. Сравнивая компоненты дисперсии друг с другом посредством F-критерия Фишера, можно определить, какая доля общей вариативности результативного признака обусловлена действием регулируемых факторов [7].

ЭТАП 3 РЕЗУЛЬТАТЫ И НЕОБХОДИМЫЕ МЕРЫ

Последним этапом при проведении количественного анализа является визуализация полученных данных и их представление. Данный этап является очень важным, хотя многие забывают про него и считают, что представление результатов менее важно, чем сам результат. Но если есть человек, который будет оценивать результаты вашего анализа и ориентироваться на них, а обычно это ваш руководитель, то он будет смотреть на ваше представление результата, а не на голые цифры.

Визуализация данных — это наглядное представление массивов различной информации.

Визуализация данных находит широкое применение большом количестве сфер общества, таких как:

- научные исследования;
- новостные сводки;
- обучение;
- аналитических обзорах и др [8].

Выделяют различные типы визуализации:

- Схематичные формы представления информации. К ним относят различные диаграммы и графики.
- Формы, усиливающие восприятие информации. Такие как диаграмма Эйлера или карта и полярный график.
- Стратегическая визуализация, которая служит для представления в визуальной форме многих аспектов деятельности организаций.
- Концептуальная визуализация, например, диаграмма Ганта или концептуальные карты.
- Комбинированная, позволяющая объединять несколько сложных графиков в один для удобства сопоставления. ярким примером данного типа визуализации может служить карта прогноза погоды.

III. Заключение

В данной статье произведен обзор аналитического подхода к проведению количественного анализа. Данный вид анализа является качественным инструментом, с помощью которого многие крупные и более мелкие компании могут оптимизировать свою деятельность, а также

Мы разработали новые теоретические основы для синтеза нелинейных фильтров, основанных на различных версиях среднего по Колмогорову. Главная цель работы – показать, что агрегационные фильтры Колмогорова могут быть использованы для решения проблем фильтрации изображений в естественной и эффективной манере.

Работа была поддержана грантами РФФИ № 13-07-12168, РФФИ № 13-07-00785 и грантом МОН РФ t №218-03-167 (согласно постановлению МОН РФ № 02.G25.31.0055 от 12.02. 2013).

IV. Литература

[1] Бородачев С.М. Имитационное моделирование в экономике [Текст]: учеб. пособие / С.М.Бородачёв. – Екатеринбург: УрФУ, 2012. – 86 с.

[2] Ким Дж.-О., Мьюллер Ч. У. «Факторный анализ: статистические методы и практические вопросы» [Текст] / сборник работ «Факторный, дискриминантный и кластерный анализ»: пер. с англ.; Под. ред. И. С. Енюкова. — М.: «Финансы и статистика», 1989. — 215 с.

[3] Кобзарь А. И. Прикладная математическая статистика. [Текст] Справочник для инженеров и научных работников. — М.: Физматлит, 2006.

[4] О чем говорят цифры. Как понимать и использовать данные [Текст] / Томас Дэвенпорт, Ким Джин Хо: Манн, Иванов и Фербер; Москва; 2014. 191 с.

[5] Яу Н. Искусство визуализации в бизнесе. Как представить сложную информацию простыми образами [Текст]: Нейтан Яу; пер. с англ. Светланы Кировой. ¬— ¬ М.: Манн, Иванов и Фербер, 2013. — 352 с.

[6] Starfield A., Smith K., and Bleloch A. How to Model It: Problem Solving for the Computer Age. [Текст] – New York: McGraw-Hill, 1994. P. 19.

[7] Дисперсионный анализ [Электронный ресурс] // URL: http://www.machinelearning.ru/wiki/index.php?title=Дисперсионный_анализ (дата обращения: 29.11.15)

[8] Зачем и как использовать визуализацию данных? [Электронный ресурс] // URL: http://habrahabr.ru/company/devexpress/blog/240325

[9] Кластерный анализ [Электронный ресурс] // URL: http://chapters.comsoc.org/tomsk/system/cluster.html

[10] Проверка гипотез [Электронный ресурс] // URL: http://statistica.ru/theory/proverka-gipotez/

[11] Data Mining [Электронный курс] // URL: http://www.intuit.ru/studies/courses/6/6/lecture/158?page=2